DOCUMENT RESUME

ED 150 510 CG 012 177

AUTHOR Leventhal, Les; And Others
TITLE Bogus Evidence for the Validity of Student Ratings.
PUB DATE Aug 77
NOTE 24p.; Paper presented at the Annual Convention of the American Psychological Association (San Francisco, California, August 26-30, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS College Students; *Course Evaluation; *Faculty Evaluation; Higher Education; *Research Methodology; Research Projects; Statistical Analysis; *Student Evaluation of Teacher Performance; *Validity

ABSTRACT

Researchers often claim validity for student ratings after finding a positive ratings/achievement correlation. This allegedly shows that ratings identify teachers who promote learning. But if students are not randomly assigned to classes, or if the correlation is not computed across classmeans, the correlation may result instead from a student characteristic that: (1) varies across classes, (2) correlates with ratings, and (3) correlates with achievement. We asked undergraduates (N=405) in a multi-section course about importance of teaching ability or reputation to their section selection. We found importance of ability/reputation (1) varied across sections, (2) correlated with end-of-course ratings of teachers, and (3) correlated with college GPA and course grade. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED150510

Bogus Evidence for the Validity of Student Ratings

Les Leventhal, Philip C. Abrami, and Raymond P. Perry

University of Manitoba

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Les Leventhal

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented at the meeting of the American Psychological Association, San Francisco, August 1977.

EC012177

Abstract

Researchers often claim validity for student ratings after finding a positive ratings/achievement correlation. This allegedly shows that ratings identify teachers who promote learning. But if students are not randomly assigned to classes, or if the correlation is not computed across classmeans, the correlation may result instead from a student characteristic that (a) varies across classes, (b) correlates with ratings, and (c) correlates with achievement. We asked undergraduates (N = 405) in a multi-section course about importance of teaching ability or reputation to their section selection. We found importance of ability/reputation (a) varied across sections, (b) correlated with end-of-course ratings of teachers, and (c) correlated with college GPA and course grade.

Bogus Evidence for the Validity of Student Ratings

College teachers are often evaluated by students responding to a teacher rating form (TRF). These ratings have become the subject of an enormous research effort. See, for example, the excellent reviews by Doyle (1975) and Feldman (1976).

Many researchers try to validate ratings in a multi-section course by measuring the correlation between ratings and student achievement (Abrami, Leventhal, & Perry, Note 1). A significant correlation is interpreted as evidence that ratings predict teacher-produced achievement. This interpretation is reasonable provided that students were randomly assigned to sections and that the correlation was computed across class-means rather than across students within a class or students pooled from all classes (Leventhal, 1975). Failure to randomize students or failure to use classmeans as the unit of analysis will permit a ratings/achievement correlation to arise from a variety of causes having nothing to do with whether teacher differences affect student achievement or ratings, and having nothing to do with whether ratings predict teacher-produced achievement (Leventhal, Perry, & Abrami, 1977). For example, a ratings/achievement correlation computed across <u>classmeans</u> in an <u>unrandomized</u> design will result from any student characteristic, such as interest in quality instruction, provided that it (a) varies from class to class, (b) correlates with individual differences among students in the way they rate teachers, and (c) correlates with individual differences among students in their ability to achieve. Furthermore, a ratings/achievement

correlation in a randomized design, computed either within a class or across pooled students, will result from any student characteristic, such as interest in quality instruction, provided that it (a) varies within each class, (b) correlates with student differences in TRF ratings, and (c) correlates with student differences in achievement.

These are tricky methodological issues and will be amplified in the discussion section. There they can be integrated with the results of our investigation into a student characteristic that relates to ratings and achievement. The importance of these issues is demonstrated in Feldman's (1976) comprehensive review in which out of hundreds of studies investigating the ratings/achievement relationship only a few have both randomized students and used classmeans as the unit of analysis.

While researchers frequently use the classmean as the unit of analysis, they rarely randomize students. Hence, discovery of a student characteristic that varies across sections and correlates with ratings and achievement will provide a basis for reinterpreting prior unrandomized studies whose ratings/achievement correlations were thought to show that ratings predict teacher-produced achievement. Furthermore, assuming that the characteristic which correlates with ratings and achievement also varies within each section will provide a basis for reinterpreting prior randomized and unrandomized studies which computed the ratings/achievement correlation either within a class or across pooled students.

Accordingly, we investigated a multi-section Introductory Psychology course at the University of Manitoba, trying to find such a student characteristic. We questionned students at the beginning of the course on biographical

items and reasons for selecting their section. At the end of the course we asked the students to rate their instructors on a TRF. Finally, we attempted to obtain the students' college GPA's and course grades from the university registrar. Since the course lasted the entire academic year, and since it required another two years to obtain GPA's and grades from the registrar(!), we reported the results of the study in installments as data became available. We found that some student characteristics varied significantly from section to section (Leventhal, Abrami, Perry & Breen, 1975). For example, Introductory Psychology sections varied in age, sex, high school grade point average, and hometown population. Another characteristic was student interest in instructor's teaching ability and/or reputation as a teacher. It was found that 21% of the students in a first year psychology course and 39% of the students in an advanced psychology course (both multi-section courses) claimed to base their section choice mostly or entirely on teaching ability/reputation. Furthermore, the fact that significant section to section difference occurred in each course in the number of students claiming this criterion to be important indicated that students actually selectively registered in sections according to it. Leventhal, Abrami, and Perry (1976) reported that, within each section, students who claimed teaching ability/reputation important to section selection rated their instructors more favorably than students who claimed teaching ability/reputation to be less important. Thus, interest in quality instruction (a) varied from section to section and (b) correlated with ratings. The present paper reports the results of the final stage of the investigation: whether interest in quality instruction correlated with achievement. Achievement was measured by final course grade and by college GPA.

Method

Subjects and Setting

The subjects were 940 students enrolled in 13 sections of a two-semester Introductory psychology course during the academic year 1973-1974. Participation in the study was voluntary for all students. During the year prior to this study, the University of Manitoba Student Union (UMSU) constructed, administered, and analyzed a TRF for all departments in the Faculty of Arts, which included Psychology. The results of these ratings were published and made widely available at the time students registered for the course investigated in this study.

Materials

To investigate the section selection process, a 22-item questionnaire was constructed relating to personal history, reasons for section choice, and sources of information used by the student for secion choice. The questionnaire is described in detail by Leventhal et al. (1975). The item of special interest to the present study asked, "How much was the selection of this section based upon your knowledge of your instructor's teaching ability and/or reputation as a teacher?" Response alternatives to this item ranged on the following four-point scale: selection based "entirely" upon (one point), "mostly" upon (two points), "somewhat" upon (three points), and "not at all" upon (four points) the reason. An "I don't remember" alternative was also provided.

Procedure

The section selection questionnaire was administered during the third and fourth week of the first semester. Students were asked in class to respond to this questionnaire to help the investigators study "why students

enroll in a particular section of a multi-sectoin course." Answers were guaranteed to be treated as confidential. Students were asked to provide their name and student numbers so that their responses to the questionnaire could be matched to their responses to another questionnaire to be administered later in the term. About two weeks prior to the end of the second semester, the TRF was administered. About two years later, college GPA's and course grades were obtained from the registrar. The GPA was computed only for the courses taken during the year in which we studied the Introductory Psychology course and includes Introductory Psychology.

Results

Only 9 of the 13 sections responding to the section selection questionnaire were available at the end of the course to respond to the TRF. In addition, students who failed to respond to either questionnaire, or answered "I don't remember" to the teacher reputation item, were dropped from analysis. As a result, the relationship between interest in quality instruction and TRF ratings reported in Leventhal et al. (1976) was based on 417 students. College GPA's and final grades for 405 of those students were obtained from the registrar and were used in the following analysis of the relationship between interest in quality instruction and achievement.

To determine whether students interested in quality instruction achieved differently than classmates, a multivariate analysis of variance (MANOVA) was performed using Finn's (1976) multivariance program where response to the ability/reputation item from the section selection questionnaire was one independent variable and sections was the other. College GPA and course grade were the two dependent variables. This analysis showed that students selecting teachers on the basis of teaching ability or reputation

achieved differently than his/her classmates, multivariate $F(6,738) = 4.01$, $p < .001$. In addition, achievement significantly varied from section to section, multivariate $F(16,738) = 3.63$, $p < .001$. Finally, the nonsignificant interaction indicated that the relationship between responses to the ability/reputation item and achievement did not significantly vary from section to section, multivariate $F(46,738) = 1.16$, $p < .22$.

Since unequal numbers of students selected the different alternatives of the ability/reputation item and since sections varied in enrollment, the number of observations per MANOVA cell in the above analysis were unequal, indicating statistically correlated independent variables. To solve this problem, Finn's multivariance program employs the a priori ordering approach to least squares analysis, described by Overall and Spiegel (1969) as Method III. This technique requires the investigator to order the independent variables in a causal or theoretical sequence which determines the order the variables were tested in the MANOVA. In the above MANOVA, and in subsequent univariate analyses, ability/reputation was tested first, sections second, and the interaction last. This order was selected because of evidence in Leventhal et al. (1975) that students selectively registered in sections according to the importance of teacher's ability/reputation; therefore, ability/reputation not only played a role prior to the effect of sections but also, in part, determined section selection.

Univariate F ratios were computed for both dependent variables, and are presented in Table 1. These analyses showed that students who selected

---

Insert Table 1 about here

---

their teacher on ability/reputation earned significantly different college GPA's and significantly different course grades than classmates using other criteria. Univariate analysis showed that significant section-to-section differences occurred for course grade, but not for college GPA. Finally, no Ability/Reputation x Section interaction occurred for either college GPA or course grade. The absence of an interaction indicates that both relationships--(a) between response to the ability/reputation item and college GPA and (b) between response to the ability/reputation item and course grade--also occur for the students within any individual section.

GPA's and course grades for each response category of the ability/reputation item were averaged, collapsing across sections. This amounts to pooling students from all sections and sorting them into groups according to their response to the ability/reputation item. The means for these response groups appear in Table 2.

Insert Table 2 about here

Inspection of GPA and grade means in Table 2 shows that, for both measures, the more a student claimed to use ability/reputation to select his/her teacher, the generally higher was the student's achievement. For both measures an exception occurred where achievement for students responding "entirely" to the ability/reputation item was lower than for students

responding "mostly." We believe that whether the population means for "entirely" and "mostly" are inverted is not as important as whether the population means as a whole can be fairly described by stating that students who used ability/reputation to select their teacher generally earned higher grades than students less interested in that criterion. To further test this statement, we performed two-tailed t tests for unequal groups on the two measures (Kirk, 1968, p.74). We grouped together students responding "entirely" or "mostly" and found that they earned significantly higher GPA's than students responding "somewhat" or "not at all," t (370) = 2.21, $p < .027$. Similarly, we found students responding "entirely" or "mostly" earned significantly higher course grades than students responding "somewhat" or "not at all," t (370) = 3.04, $p < .003$.

Discussion

The results indicated that students who selected their section on the basis of the teacher's ability or reputation generally earned higher grades than classmates less concerned with ability/reputation. This finding has implications for TRF field use and for those who try to validate ratings by using rating/achievement correlations to conclude that ratings predict teacher-produced achievement.

Field Use of TRF's

TRF ratings collected in the field frequently are published to help students select instructors. Our findings suggest that students who use ability/reputation to select instructors also earn higher grades than classmates. The result may be a stable achievment cycle in which instructors with favorable published TRF ratings attract students who later earn higher grades which, in turn, may contribute favorably to the teacher's reputation.

There are many mechanisms by which high achieving students may contribute to their teacher's reputation—ranging from departmental reviews of instructor effectiveness to a direct cause-effect link in which students who earn high grades award high ratings in gratitude. While some of these mechanisms have been investigated, the research is not yet definitive. The stable achievement cycle is best regarded as a possiblity.

More important than a stable achievement cycle is the possibility that high student achievement may be interpreted by the instructor as evidence of his/her teaching ability rather than a consequence of previously published ratings. Similarly, low achievement may be interpreted as evidence of lack of ability. Thus, some teachers may retain techniques that should be replaced, and others may replace techniques that should be retained.

TRF Validation

Our finding that students selecting their section on ability/reputation earn higher grades than classmates, together with the findings that (a) students selecting their section on ability/reputation rated their teacher more favorably than classmates (Leventhal et al., 1976),[1] and (b) student interest in quality instruction varied significantly from section to section (Leventhal et al., 1975), suggest that at least one student characteristic found to vary across sections also related to the way students rate teachers and to student achievement. These findings have implications for the assignment of students to sections and the unit of analysis employed when a ratings/achievement correlation is used as evidence that ratings predict teacher-produced achievement.

Assignment of students to sections. The findings suggest (a) that one

mechanism was at work at the University of Manitoba which would promote a positive ratings/achievement correlation across classmeans in a non-random design, and (b) that this mechanism will produce the correlation even when teachers affect neither ratings nor achievement and when ratings do not predict teacher-produced achievement. The importance of random assignment of students to sections is clear: randomization will render improbable section-to-section variation in student characteristics. Hence, student characteristics that correlate with both ratings and achievement are held equivalent across sections and cannot produce a ratings/achievement correlation across classmeans.

Unit of analysis. While the implication of the present findings for nonrandom designs are straightforward, the implications for the unit of analysis employed in the computation of the ratings/achievement correlation are more subtle. Researchers have used three different units of analysis. The ratings/achievement correlation has been computed (a) across students within a single class and averaged with correlations from the remaining classes, (b) across classmeans, and (c) across all students pooled together regardless of class. Researchers have disagreed on the proper unit of analysis. Even researchers who agree on statistical properties of the units disagree on what causal processes affect them. For example, compare Rodin's analysis of Frey's within-class correlation to Frey's (Rodin, 1975; Frey, 1973; and Frey, 1975). Nevertheless, we believe that these causal processes must be examined to show the importance of our findings to the selection of a unit of analysis.

The key to seeing what causal processes affect the three kinds of ratings/achievement correlations is in first considering ratings

and achievement separately. Consider, for example, how student, teacher, and situational factors (e.g., classtime, level of course, workload, etc.) affect[2] (a) within-class achievement variation, (b) class-to-class variation in achievement means, and (c) variation in achievement for students pooled from all classes. In what follows we will try to suggest that variation in achievement classmeans comes closer than variation in the other units to reflecting differences in teaching ability and random error. For this reason, a ratings/achievement correlation computed across classmeans comes closer than the other correlations to identifying whether ratings predict teacher-produced achievement.

Within-class variation in achievement cannot result from differences in teaching ability or class-to-class variation in situational factors because all students within a class are exposed to the same teacher and the same situational factors. Instead, within-class variation probably reflects student-to-student variation in achievement-related characteristics such as initial ability, motivation, study habits, family background, ability to elicit special help from the instructor, etc. Anything that correlates with within-class achievement variation probably predicts the impact of these student factors on achievement. Thus, the ratings/achievement correlation computed within a class provides evidence on whether ratings predict student-produced achievement. Such a correlation can come about, for example, if students in a class who are interested in quality instruction award higher ratings and earn higher grades than less interested classmates. Thus, assuming students vary within a class in interest in quality instruction, our series of studies suggest how a within-class ratings/achievement correlation may come about, even if

students are randomly assigned to classes and teachers have no effect on achievement. Accordingly, researchers using a within-class ratings/achievement correlation to prove that ratings predict teacher-produced achievement have used ambiguous evidence.

Variation in achievement means from class to class may result from class-to-class differences in students, teachers, and situational factors. Randomly assigning students to classes will suppress the student effect. Some situational effects can be reduced by studying multiple sections of the same course which have common workloads, classtimes, examinations, etc. Using these techniques, class-to-class variation in achievement means can be made, more or less, to reflect only teacher-produced achievement and random error. Thus, the ratings/achievement correlation computed across classmeans provides evidence on whether ratings predict teacher-produced achievement, provided that students were randomized and that situational factors were controlled. Without randomization and situational control, a correlation across classmeans may come about, for example, if students interested in quality instruction rate teachers well, earn high grades, and seek out highly reputed teachers. Here, ratings have predicted student-produced achievement. Thus, the results of our series of studies suggest how a ratings/achievement correlation across classmeans may come about, even if teachers have no effect on achievement.

Variation in achievement for students pooled from all classes will be produced by whatever produces within-class achievement variation and whatever produces variation in achievement classmeans. Accordingly, achievement variation for pooled students, even when students are randomly assigned to classes, will reflect student-, teacher-, and situationally-produced

achievement. Anything that correlates with pooled achievement variation probably predicts achievement that is produced by all factors combined. Thus the ratings/achievement correlation that is computed across pooled students probably provides evidence on whether ratings predict achievement produced by all factors combined. In this way, the pooled ratings/achievement correlation found by Kennedy (1975) in a randomized design can be explained by anything that explains, say, a within-class correlation; for example, students interested in quality instruction vary within classes, earn higher grades, and award higher ratings than interested classmates. Thus, our series of studies suggest how a ratings/achievement correlation across students pooled from all classes may come about, even if teachers have no effect on achievement.

We have tried to show the importance of student characteristics that correlate with ratings and achievement to the selection of a unit of analysis for the ratings/achievement correlation. We have not attempted to use our findings to prove which unit is correct--that is a methodological, not empirical, issue. Nor do we believe our findings to be ubiquitous phenomena that always produce ratings/achievement correlations when specific methodlogies are employed--even though we continually invoked our findings to illustrate reinterpretations of those correlations. Our point is that all ratings/achievement correlations arising from some research methodologies can be attributed to any student characteristic that correlates with ratings and achievement and, furthermore, one such student characteristic was found operating at the University of Manitoba.

We do not believe the results of this study should be used to reject the notion that TRF's can predict teacher-produced achievement. Indeed, the positive ratings/achievement correlations found by Sullivan and Skanes (1974) and Centra (1977) cannot be reinterpreted by our findings (because they computed the correlation across classmeans in a randomized design). Instead, we believe that one can determine whether ratings identify teachers who promote learning. But our findings suggest that investigators must attend more to the design and analysis of the research than has commonly been the case.

Cautions

These findings should be interpreted with the cautions described in Leventhal et al. (1975) and Leventhal et al. (1976). Three other cautions are in order.

First, our conclusion that students selecting their section because of their teacher's ability or reputation rested on finding a significant main effect of the ability/reputation item coupled with the absence of a significant ability/reputation x section interaction. A significant interaction would have indicated that the relation between response to the ability/reputation item and achievement changed from section to section, perhaps enough to void our conclusion. Unfortunately, our statistical test for the interaction had little power; hence, our ability to find it was limited. The lack of power resulted from our a priori decision to test the interaction after testing the two main effects.

Second, the design of our study does not permit determining whether interest in quality instruction is simply correlated with achievement or instead caused the high grades. (High grades may result, for example, when a student works harder because he/she thinks the teacher is highly reputed.) Nevertheless, little of the discussion of TRF field use or validation depends on the relation being correlational or causal.

Third, the reader should bear in mind that the three research reports --Leventhal et al. (1975), Leventhal et al. (1976), and the present one-- constitute a single study whose duration prompted separate reports as data became available. Consequently, the results should not be treated as independent findings that wonderfully mesh together. Instead, they are prone to similar errors, idiosyncracies, and prejudices that result from the "three" studies being conducted on the same students by the same researchers in the same setting.

Reference Note

1. Abrami, P., Leventhal, L., & Perry, R. The generalizability of student ratings of instruction. Paper presented at the meeting of the American Educational Research Association, New York, April 1977.

References

Centra, J. Student ratings of instruction and their relationship to student learning. American Educational Research Journal, 1977, 14, 17-24.

Doyle, K. Student evaluation of instruction. Lexington: D.C. Heath, 1975.

Feldman, K. Grades and college student's evaluations of their courses and teachers. Research in Higher Education, 1976, 4, 69-111.

Finn, J. Multivariance-Univariate and multivariate analysis of variance and covariance, and regression (Version V). Chicago: National Educational Resources, 1976.

Frey, P. Student ratings of teaching: Validity of several rating factors. Science, 1973, 182, 83-85.

Frey, P. Untitled contribution to a section entitled: Student evaluation. Science, 1975, 187, 555-559.

Kennedy, W. Grades expected and grades received--their relationship to students' evaluations of faculty performance. Journal of Educational Psychology, 1975, 67, 109-115.

Kirk, R. Experimental design: Procedures for the behavioral sciences. Belmont: Brooks/Cole, 1968.

Leventhal, L. Teacher rating forms: Critique and reformulation of previous validation designs. Canadian Psychological Review, 1975, 16, 269-276.

Leventhal L., Abrami, P., Perry, R., & Breen, L. Section selection in multisection courses: Implications for the validation and use of teacher rating forms. Educational and Psychological Measurement, 1975, 35, 885-895.

Leventhal, L., Abrami, P., & Perry, R. Do teacher rating forms reveal as much about students as about teachers? Journal of Educational Psychology, 1976, 68, 441-445.

Leventhal, L., Perry, R., & Abrami, P. Effects of lecturer quality and student perception of lecturer's experience on teacher ratings and student achievement. Journal of Educational Psychology, 1977, 69, 360-374.

Overall, J., & Spiegel, D. Concerning least squares analysis of experimental data. Psychological Bulletin, 1969, 72, 311-322.

Rodin, M. Untitled contribution to a section entitled: Student evaluation. Science, 1975, 187, 555-559.

Sullivan, A., & Skanes, G. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 584-590.

Footnotes

This research was supported by grants from the University of Manitoba Research Board, the University of Manitoba Group for Advance in Innovative Instruction, and the Canada Council (S76-0345).

Portions of this article were presented at the meeting of the American Psychological Association, San Francisco, August 1977.

Requests for reprints should be sent to Les Leventhal, Department of Psychology, University of Manitoba, Winnipeg, Canada, R3T 2N2.

[1]An inversion occurred in Leventhal et al. (1976) similar to the "entirely"/"mostly" inversion reported here. Leventhal et al., studying the same students studied here, found that students responding "entirely" to the reputation/ability item rated their instructor less favorably on an overall teaching ability item than students responding "mostly." We reanalyzed these old data in order to compare students responding "entirely" or "mostly" to students responding "somewhat" or "not at all," as was done in the present study. Using a two-tailed t test for unequal groups (Kirk, 1968, p. 74), we found that students responding "entirely" or "mostly" to the reputation/ability item rated their instructor more favorably than students responding "somewhat" or "not at all," t (382) = 2.50, $p < .013$.

[2]For simplicity, main effects of these factors will be considered--not interactions among them.

Table 1

Univariate Analysis of Variance for College GPA's and for Course Grades

College GPA				Course Grade			
Source	df[a]	MS	F	Source	df[a]	MS	F
Ability/ Reputation (A)	3	3.51	6.60**	Ability/ Reputation (A)	3	4.92	6.18**
Sections (B)	8	0.47	0.89	Sections (B)	8	2.16	2.71*
A x B	23	0.69	1.31	A x B	23	0.81	1.02
Error	370	0.53		Error	370	0.80	

[a] There was one ANOVA cell with no observations, resulting in the loss of a degree of freedom for the A x B interaction.

*p .01

**p .001

Table 2

Means for College GPA's and for Course Grades

Importance of Ability/Reputation	Mean College GPA[a]	Mean Course Grade[b]
Entirely	2.89	3.22
Mostly	2.96	3.31
Somewhat	2.88	3.05
Not at all	2.59	2.83

[a]GPA's were obtained from the university registrar. To compute a GPA, the registrar assigned points to course grades as follows: A+ or A = 4, B+ = 3.5, B = 3, C+ = 2.5, C = 2, D = 1, F = 0.

[b]Letter grades for the Introductory Psychology course were obtained from the registrar. We assigned points to the letter grades as follows: A+ = 4.5, A = 4, B+ = 3.5, B = 3, C+ = 2.5, C = 2, D = 1, F = 0. The preceding letter grades are the only ones used in courses that are not pass/fail.